Dynamic classification using credible intervals in longitudinal discriminant analysis

Journal:	*Statistics in Medicine*
Manuscript ID	SIM-16-0899.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Hughes, David; University of Liverpool, Biostatistics Komárek, Arnošt; Charles University, Probability and Statistics Bonnett, Laura; University of Liverpool, Biostatistics Czanner, Gabriela; University of Liverpool, Biostatistics; University of Liverpool, Eye and Vision Science Garcia-Finana, Marta; University of Liverpool, Biostatistics
Keywords:	Longitudinal Discriminant Analysis, Credible intervals, Allocation Scheme, Multivariate Classification, Biomarkers

Statistics in Medicine

Research Article

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Dynamic classification using credible intervals in longitudinal discriminant analysis

David M. Hughes[a]*, Arnošt Komárek[b], Laura J. Bonnett[a], Gabriela Czanner[a,c] and Marta García-Fiñana[a]

Recently developed methods of Longitudinal Discriminant Analysis allow for classification of subjects into prespecified prognostic groups using longitudinal history of both continuous and discrete biomarkers. The classification utilises Bayesian estimates of the group membership probabilities for each prognostic group. These estimates are derived from a multivariate generalized linear mixed model of the biomarker's longitudinal evolution in each of the groups, and can be updated each time new data is available for a patient, providing a dynamic (over time) allocation scheme. However, the precision of the estimated group probabilities differs for each patient and also over time. This precision can be assessed by looking at credible intervals for the group membership probabilities. In this paper, we propose a new allocation rule that incorporates credible intervals for use in context of a dynamic longitudinal discriminant analysis, and show that this can decrease the number of false positives in a prognostic test, improving the Positive Predictive Value (PPV). We also establish that by leaving some patients unclassified for a certain period of time, the classification accuracy of those patients who are classified can be improved, giving increased confidence to clinicians in their decision making. Finally, we show that determining a stopping rule dynamically can be more accurate than specifying a set time point at which to decide on a patient's status. We illustrate our methodology using data from patients with epilepsy and show how patients who fail to achieve adequate seizure control are more accurately identified using credible intervals compared to existing methods.

Keywords: Credible intervals; allocation scheme; longitudinal discriminant analysis.

1. Introduction

In many medical studies, measurements from patients are regularly taken over time on multiple clinical markers. Interest may be in the future prognosis of individual patients, e.g., whether a patient will suffer renal graft failure within ten years of transplant [1]. In this paper we focus on dynamic longitudinal discriminant analysis (LoDA) where patients are classified into one of several prognostic groups (based on future status) using their clinical history, and the classification is updated each time new data becomes available.

[a] *Department of Biostatistics, University of Liverpool,UK*
[b] *Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic*
[c] *Department of Eye and Vision Science, University of Liverpool, UK*
* *Correspondence to: Department of Biostatistics, Block F, Waterhouse Building, 1–5 Brownlow Street University of Liverpool, Liverpool, L69 3GL,UK Email:dmhughes@liverpool.ac.uk*

Prepared using simauth.cls *[Version: 2010/03/10 v3.00]*

Statistics in Medicine

LoDA has been developed by a number of authors over relatively recent history. Initial interest focused on using a single continuous longitudinal marker to predict group membership for a patient [2, 3, 4, 5, 6]. Multivariate LoDA with continuous markers have also been considered [7, 8, 9]. However, some of the longitudinal markers may not be continuous, but rather binary or counts. Fieuws *et al.* [1] presented a multivariate LoDA method for both continuous and binary longitudinal markers. An alternative multivariate LoDA method for longitudinal markers of different types (continuous, counts, binary), which is robustified against possible model misspecification, was recently developed by Hughes *et al.* [10]. These models use the longitudinal history of patients of known prognosis to develop a classification procedure that can be used to classify new patients based on their own longitudinal data. For each patient a probability of belonging to each prognostic group is calculated and used in an allocation scheme to assign the new patient to a group. These probabilities are re-evaluated for a patient each time new clinical data is available, hence we refer to dynamic LoDA.

In this paper, we aim to further improve the approach of Hughes *et al.* [10] who applied their method to data from the Standard and New Antiepileptic Drugs (SANAD) study (Marson *et al.* [11, 12]), a randomised control trial comparing treatments for patients with newly diagnosed epilepsy. Patients were recruited between December 1999 and August 2004, and underwent regular surveillance. At recruitment, baseline information was collected including the type of epilepsy a patient had, their age and sex. At regular follow up visits, information was collected regarding the treatment they received, how many seizures the patient had experienced since their previous visit and what adverse events had been observed. Patients were followed up until potentially January 2006.

The approach of Hughes *et al.* [10] aimed to identify those patients who will fail to achieve remission from seizures within a 5-year follow-up period. Patients who achieve a continuous 12 month period free from seizures within five years of diagnosis are regarded as being in *remission*, whereas patients who do not are referred to as *refractory*. Clinical interest is in being able to assess, each time a patient attends follow up, their risk of ultimately belonging to the refractory group. In this approach, a patient's group membership (remission/refractory) does not change and is based on their observed condition at five years from initial diagnosis. Good levels of classification accuracy were achieved by Hughes *et al.* [10], with sensitivity and specificity showing values above 90%. However, a positive predictive value (PPV) of 59% was reported, which implies that 41% of patients who were classified using the LoDA approach as not achieving remission of seizures, did in fact achieve remission. This low PPV was influenced by the relatively low incidence value (only 10% of all the patients were refractory). The clinical implications of wrongly classifying someone as refractory are important (e.g., surgery could be considered for these patients). It is hence desirable that a prognostic test performs well, not only in terms of sensitivity and specificity, but also in terms of the PPV. In this paper, we propose different allocation schemes which account for patient specific variability and improve the PPV whilst maintaining good levels of specificity and sensitivity.

To formalize our setting, we consider a situation in which regular measurements are made over time (i.e., longitudinally) of $R \geq 1$ markers. For each patient, one of G prespecified diagnoses is made at a specific future time T. This will be represented by a value of the random variable $U \in \{0, \ldots, G-1\}$, observable only at time T. In this way, the population of patients is split into G groups depending on their future status at time T. In the SANAD example, $G = 2$, with $U = 0$ denoting the remission group and $U = 1$ denoting the refractory group, with T being defined as 5 years from initial diagnosis. We denote all the longitudinal observations of a particular marker for a particular patient by $\mathbf{Y}_r = \left(Y_{r,1}, \ldots, Y_{r,n_r}\right), r = 1, \ldots, R$ where the observations have been (are to be) recorded at times $\mathbf{t}_r = \left(t_{r,1}, \ldots, t_{r,n_r}\right)$, $t_{r,1} < \cdots < t_{r,n_r} < T$, possibly with additional covariate vectors $\mathbf{v}_{r,1}, \ldots, \mathbf{v}_{r,n_r} \in \mathbb{R}^{p_r}$ overall denoted as $\mathcal{C}$. Our aim is to utilise the information collected about relevant markers to predict the future group, U, to which a patient belongs. We aim to do this dynamically, by which we mean that we update our prediction of a patient's future prognosis each time we have new available marker and covariate data for that patient. In other words, for a given t, $0 < t < T$, let $\mathbf{Y}_r(t) = \left(Y_{r,j} : t_{r,j} \leq t\right)$, $r = 1, \ldots, R$, $\mathbb{Y}(t) = \left(\mathbf{Y}_1(t), \ldots, \mathbf{Y}_R(t)\right)$ denote the values of the longitudinal markers gathered by time t. Similarly, let $\mathcal{C}(t)$ denote the covariate information by time t. Each time a patient visits a clinic for follow up (at time t), marker and covariate information is collected leading to $\mathbb{Y}(t)$ and $\mathcal{C}(t)$ and the prediction $\widehat{U}(t)$ of the patient's future prognosis is calculated based on available data $\mathbb{Y}(t)$ and $\mathcal{C}(t)$. That is, the prediction $\widehat{U}(t)$ exploits both the

newly collected data at time t as well as all previously gathered data for that patient.

To calculate the predictions $\widehat{U}(t)$, $t < T$, we assume that historical data (denoted as $\mathcal{Y}$) is available for patients whose longitudinal marker and covariate history and also their prognosis (group membership) are known. This is used along with the marker and covariate history $\mathbb{Y}(t)$ and $\mathcal{C}(t)$ of a particular (new) patient to calculate estimates $\widehat{\mathcal{P}}_g(t)$, $g = 0, \ldots, G-1$ of the group membership probabilities $\mathcal{P}_g(t) = \mathsf{P}\big(U = g \,\big|\, \mathbb{Y}(t), \mathcal{C}(t), \mathcal{Y}\big)$ for this patient. The estimated group membership probabilities $\widehat{\mathcal{P}}_0(t), \ldots, \widehat{\mathcal{P}}_{G-1}(t)$ are then used to determine the prediction $\widehat{U}(t)$ by applying a suitable allocation scheme. Classically, $\widehat{U}(t) = \mathsf{argmax}_{g=0,\ldots,G-1}\, \widehat{\mathcal{P}}_g(t)$, but other schemes are possible as well (see Section 2.4).

In all of the LoDA procedures referenced above ([1]–[10]), point estimates $\widehat{\mathcal{P}}_g(t)$, $g = 0, \ldots, G-1$, of the group membership probabilities are used to derive the prediction $\widehat{U}(t)$, i.e., to assign a new patient to a group. However, the precision of these estimated probabilities may be different for different patients and also at different occasions. This is due to the statistical error of the estimation procedure which relates mainly to different amount of information borne by the longitudinal marker and covariate information $\mathbb{Y}(t)$ and $\mathcal{C}(t)$ for different patients and different time points t. Indeed, for each individual patient, the amount of useful information to predict the group membership increases with t while the statistical error of the estimated group membership probabilities decreases. This statistical error can be assessed by using credible/confidence intervals (depending on whether a frequentist or Bayesian methodology is used to fit the underlying models) around the estimated group membership probabilities. We aim to incorporate this additional information in a classification scheme and, thus reduce the number of false positives and false negatives obtained. The idea is to leave a small number of patients unclassified for whom we are least certain of their prognosis, whilst classifying patients for whom we are confident of their prognosis. We propose a new allocation scheme which takes into account the variability of the estimation of the group membership probabilities using the multivariate LoDA method of Hughes *et al.* [10] by incorporating information provided by the corresponding credible intervals.

To the best of our knowledge, this is the first time credible (or confidence) intervals for the group membership probabilities have been used in dynamic LoDA, although we are not the first to consider credible intervals in a classification scheme. Komárek and Komárková [13] do so in the context of longitudinal cluster analysis (where the classification is not 'dynamic' and clustering (i.e., unsupervised classification) is performed only once using all available longitudinal data) whilst Gugliemli *et al.* [14] do so in the context of survival analysis (although we note that this is not in a longitudinal context, and there is no sequential updating). Leaving a group as unclassified is similar to the neutral zone classifiers proposed by Zhang *et al.* [15], although their unclassified group is based upon an analysis of the misclassification costs. Horrocks and van Den Heuvel [16] use point estimates to allocate patients to clinical groups and then use credible intervals around the point estimates to help inform patient decisions. Shah *et al.* [17] calculate a confidence interval around each group membership probability in order to assess the individual discriminative ability of each patient.

An outline of the remainder of the paper is as follows. In Section 2 we give a brief outline of a multivariate LoDA procedure recently developed by Hughes *et al.* [10] and provide a review of the most commonly used classification rules. We introduce our new classification scheme based on credible intervals in Section 3. This proposed scheme is applied to the SANAD data in Section 4 where we demonstrate its potential benefits. Section 5 compares our new scheme with the existing classification schemes described in Section 2. Finally, we give a brief discussion of our findings in Section 6.

2. LoDA based on a Multivariate Generalised Linear Mixed Model

2.1. *Group specific multivariate generalised linear mixed model*

Our proposal to classify a new patient on the basis of credible intervals for the individual group probabilities (explained in Section 3) starts from considering the following procedure developed by Hughes *et al.* [10] based on the multivariate generalised linear mixed model (MGLMM) proposed by Komárek and Komárková [13]. It is assumed that given $U = g$ (given the allocation of a particular patient into group g, $g = 0, \ldots, G-1$), the values of the longitudinal markers

Statistics in Medicine

$\mathbf{Y}_1, \ldots, \mathbf{Y}_R$ are generated by the group specific MGLMM. Namely, given $U = g$ and given a latent random effects vector $\mathbf{b} = (\mathbf{b}_1, \ldots, \mathbf{b}_R)$, the jth longitudinal observation $Y_{r,j}$ ($j = 1, \ldots, n_r$) of the rth marker ($r = 1, \ldots, R$) is assumed to follow a distribution from an exponential family with a dispersion parameter ϕ_r^g and the expectation given as

$$h_r^{-1}\Big\{\mathsf{E}\big(Y_{r,j} \,\big|\, \mathbf{b}, U = g\big)\Big\} = \mathbf{x}_{r,j}^{g\top}\boldsymbol{\alpha}_r^g + \mathbf{z}_{r,j}^{g\top}\mathbf{b}_r, \qquad r = 1, \ldots, R,\ j = 1, \ldots, n_r, \tag{1}$$

where h_r^{-1} is a chosen link function, $\mathbf{x}_{r,j}^g = \mathbf{x}_{r,j}^g(\mathcal{C})$ and $\mathbf{z}_{r,j}^g = \mathbf{z}_{r,j}^g(\mathcal{C})$ are covariate vectors used in a model for the prognostic group g derived from the marker information and the available covariates $\mathcal{C}$ known by the time at which $Y_{r,j}$ was measured. Parameter vectors $\boldsymbol{\alpha}_r^g$, $r = 1, \ldots, R$, $g = 0, \ldots, G-1$ are unknown regression coefficients.

The random effects vectors $\mathbf{b}_1, \ldots, \mathbf{b}_R$ are included in the model formula (1) to account for possible correlation between measurements of the same marker on a particular patient. To account also for the correlation between measurements of different markers on a particular patient, a joint distribution with possibly non-diagonal covariance matrix is considered for the overall random effects vector $\mathbf{b}$. Parameters of this joint distribution are again group specific. As a certain form of robustification of the model towards misspecification of this distribution, Komárek and Komárková [13] use a multivariate normal mixture here. That is, it is assumed that

$$\mathbf{b} \,\big|\, U = g \;\sim\; \sum_{k=1}^{K^g} w_k^g\, \mathcal{MVN}\big(\boldsymbol{\mu}_k^g, \mathbb{D}_k^g\big), \tag{2}$$

where $\mathcal{MVN}(\boldsymbol{\mu}, \mathbb{D})$ stands for a multivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and a covariance matrix $\mathbb{D}$ having a density denoted as $\varphi(\cdot; \boldsymbol{\mu}, \mathbb{D})$. Unknown parameters of the mixture model (2) in the prognostic group g are: the mixture weights $\mathbf{w}^g = \big(w_1^g, \ldots, w_{K^g}^g\big)$ ($0 < w_k^g < 1$, $k = 1, \ldots, K^g$, $\sum_{k=1}^{K^g} w_k^g = 1$), the mixture means $\boldsymbol{\mu}_1^g, \ldots, \boldsymbol{\mu}_{K^g}^g$ and the mixture covariance matrices $\mathbb{D}_1^g, \ldots, \mathbb{D}_{K^g}^g$. The number of mixture components, K_g, is assumed to be known.

The MGLMM for group g, $g = 0, \ldots, G-1$, defined by (1) and (2) involves conceptually two sets of unknown parameters: $\boldsymbol{\psi}^g := \big(\alpha_1^g, \ldots, \alpha_R^g, \phi_1^g, \ldots, \phi_R^g\big)$ which are the regression coefficients and dispersion parameters related to the distribution of the longitudinal response given the random effects and $\boldsymbol{\theta}^g := \big(\mathbf{w}^g, \boldsymbol{\mu}_1^g, \ldots, \boldsymbol{\mu}_{K^g}^g, \mathbb{D}_1^g, \ldots, \mathbb{D}_{K^g}^g\big)$ related to the distribution of random effects. The model induces the following group specific density of the observable vector $\big(\mathbf{Y}_1, \ldots, \mathbf{Y}_R\big)$ of the longitudinal markers

$$f_g^{marg}\big(\mathbf{y}_1, \ldots, \mathbf{y}_R; \boldsymbol{\psi}^g, \boldsymbol{\theta}^g, \mathcal{C}\big) = \int \prod_{r=1}^{R} \prod_{j=1}^{n_r} p_r\big(y_{r,j} \,\big|\, \mathbf{b}; \boldsymbol{\psi}^g, \mathcal{C}\big) \sum_{k=1}^{K^g} w_k^g\, \varphi(\mathbf{b}; \boldsymbol{\mu}_k^g, \mathbb{D}_k^g)\, \mathrm{d}\mathbf{b}. \tag{3}$$

where $p_r\big(\cdot \,\big|\, \mathbf{b}; \boldsymbol{\psi}^g, \mathcal{C}\big)$ denotes an exponential family density of the random variable $Y_{r,j}$ related to the GLMM (1).

Several studies which use discrimination based on mixed models [9, 10, 18], describe classification derived from the group probabilities which use (3) as *marginal* prediction (of a future status of a new patient). Alternatively, in this context, so called *conditional* and *random effects* predictions are discussed, each motivated by considering a different focus on the density of the observed markers. In this paper, we solely focus on the marginal prediction approach which provided the best predictive accuracy for the SANAD application in Hughes *et al.* [10]. Nevertheless, the whole methodology described in this paper can be straightforwardly applied for the conditional and random effects prediction.

2.2. Individual group probabilities

Suppose first that all model parameters $\boldsymbol{\psi} = \big(\boldsymbol{\psi}^0, \ldots, \boldsymbol{\psi}^{G-1}\big)$, $\boldsymbol{\theta} = \big(\boldsymbol{\theta}^0, \ldots, \boldsymbol{\theta}^{G-1}\big)$ are known and the task is to classify a (new) patient whose history of the longitudinal markers by time $t < T$ is $\mathbb{Y}(t) = \big(\mathbf{Y}_1(t), \ldots, \mathbf{Y}_R(t)\big)$, $\mathbf{Y}_r(t) = \mathbf{y}_r = \big(y_{r,1}, \ldots, y_{r,n_r}\big)$, $r = 1, \ldots, R$ and the covariate information known by time t is $\mathcal{C}(t)$. By Bayes theorem, we can calculate

the following conditional probabilities that the new patient belongs to each of the G prognostic groups:

$$\begin{aligned} \mathsf{P}\big(U = g \,\big|\, \mathbb{Y}(t), \mathcal{C}(t); \boldsymbol{\psi}, \boldsymbol{\theta}\big) &= \frac{\pi_g \, f_g^{marg}\big(\mathbf{y}_1, \ldots, \mathbf{y}_R; \boldsymbol{\psi}^g, \boldsymbol{\theta}^g, \mathcal{C}(t)\big)}{\sum_{\tilde{g}=0}^{G-1} \pi_{\tilde{g}} \, f_{\tilde{g}}^{marg}\big(\mathbf{y}_1, \ldots, \mathbf{y}_R; \boldsymbol{\psi}^{\tilde{g}}, \boldsymbol{\theta}^{\tilde{g}}, \mathcal{C}(t)\big)} \\ &=: \mathcal{P}_g\big(t; \boldsymbol{\psi}, \boldsymbol{\theta}\big), \qquad g = 0, \ldots, G-1, \end{aligned} \tag{4}$$

where $\pi_g = \mathsf{P}(U = g)$, $g = 0, \ldots, G-1$, are prevalences of the prognostic groups in the study population which are assumed to be known as is common in applications of discriminant analysis.

In a more common frequentist setting, historical data $\mathcal{Y}$ containing both history of the longitudinal markers, covariate information and also information on the group membership of the subjects involved are used to obtain estimates $\widehat{\boldsymbol{\psi}}$ and $\widehat{\boldsymbol{\theta}}$ of the model parameters which are then used to calculate the group probabilities of new patients. That is, classification is based on the group probabilities $\mathcal{P}_g\big(t; \widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\theta}}\big)$, $g = 0, \ldots, G-1$.

2.3. *Bayesian estimates of the individual group probabilities*

Hughes *et al.* [10] suggest to exploit Bayesian estimates of the group probabilities (4) based on the historical data $\mathcal{Y}$ for final classification. Those are given as posterior means of (4) with respect to the posterior distribution $[\boldsymbol{\psi}, \boldsymbol{\theta} \,|\, \mathcal{Y}]$ of the model parameters estimated in a Bayesian way using the historical data. That is, allocation of a new patient is based on the following group probabilities:

$$\mathcal{P}_g(t) = \mathsf{E}_{[\boldsymbol{\psi}, \boldsymbol{\theta} \,|\, \mathcal{Y}]} \mathcal{P}_g\big(t; \boldsymbol{\psi}, \boldsymbol{\theta}\big) \qquad g = 0, \ldots, G-1. \tag{5}$$

Note that the value $\mathcal{P}_g(t)$ is not only the posterior mean of $\mathcal{P}_g\big(t; \boldsymbol{\psi}, \boldsymbol{\theta}\big)$ given the historical data $\mathcal{Y}$ but can also be expressed as

$$\mathcal{P}_g(t) = \mathsf{P}\big(U = g \,\big|\, \mathbb{Y}(t), \mathcal{C}(t), \mathcal{Y}\big). \qquad g = 0, \ldots, G-1.$$

That is, the group probabilities $\mathcal{P}_0(t), \ldots, \mathcal{P}_{G-1}(t)$ express probabilities of allocation of a (new) patient into the G prognostic groups given his longitudinal information known by time t and also given the information from the historical data. Hughes *et al.* [10] approximate the group probabilities (5) by a Markov chain Monte Carlo (MCMC) method as

$$\widehat{\mathcal{P}}_g(t) = \frac{1}{M} \sum_{m=1}^{M} \mathcal{P}_g\big(t; \boldsymbol{\psi}^{(m)}, \boldsymbol{\theta}^{(m)}\big), \qquad g = 0, \ldots, G-1, \tag{6}$$

where $\big(\boldsymbol{\psi}^{(m)}, \boldsymbol{\theta}^{(m)}\big)$, $m = 1, \ldots, M$ is an MCMC sample from the posterior distribution $[\boldsymbol{\psi}, \boldsymbol{\theta} \,|\, \mathcal{Y}]$ (see Komárek and Komárková [13] for details of the MCMC procedure and also for the full specification of the Bayesian model).

2.4. *Classification rules*

The usual procedure is to assign the subject to whichever group has the largest probability. That is, with the estimated group probabilities (6) the prediction $\widehat{U}(t)$ of the future status of a (new) patient made at time t would be $\widehat{U}(t) = \mathsf{argmax}_{g=0, \ldots, G-1} \widehat{\mathcal{P}}_g(t)$. As additional observations of the new patient are acquired, the predicted classification of a given patient can be updated by repeating the discriminant analysis step with the new information. The models fitted using the historical data do not need to be recalculated, with the MCMC based Bayesian methodology, no additional MCMC samples have to be generated. Simply the estimated group membership probabilities (6) are updated, and the chosen allocation rule is reapplied.

A number of alternative classification rules have been proposed in the statistical literature [19]. Our focus now will be on the two group classification case ($G = 2$), although the methods outlined would work similarly in a multiple group discriminant analysis ($G > 2$). For clarity of exposition, we shall refer to the two groups as *disease* (labeled by $U = 1$ and

Statistics
in Medicine

also as D) and *disease free* (labeled by $U = 0$) groups. The group probabilities (4), (5) and (6), respectively, calculated at time t and related to the disease group will be indicated by subscript D.

In the following, it is assumed that prediction of the status of a (new) patient is to be obtained at visit times $0 \leq t_1 < \cdots < t_n < T$. We consider the situation where the follow-up of a particular patient is terminated/changed once it is sufficiently certain that (s)he belongs to either of the two groups. Once the clinician is, at time t, sufficiently certain that a particular patient belongs to the disease group ($\widehat{U}(t) = 1$), future prediction of the disease status of this patient may stop as appropriate clinical procedures are followed to treat the disease. Conversely, it is also common that once the clinician is sufficiently certain that a patient belongs to the non-disease group ($\widehat{U}(t) = 0$), there is perhaps no reason to continue in the follow-up of that patient, making them undertake additional time consuming, costly or uncomfortable medical examinations. However, if neither of the group allocations are certain enough, the patient remains unclassified ($\widehat{U}(t) = \mathsf{N.A.}$) and continues follow-up till either classification is determined or the disease status is known at time T.

As well as the prediction accuracy of the classification procedure, it is also of interest to evaluate the time needed to arrive at a final prediction of the patient's status. This time will be called the *prediction time* T_{pred} being defined as $T_{pred} = \min\{t : \widehat{U}(t) \neq N.A.\}$. The remaining time till the true status of the patient is known will then be called as the *lead time* T_{lead}, i.e., $T_{lead} = T - T_{pred}$. If patient remains unclassified after the last examination, we define $T_{pred} = T$ and $T_{lead} = 0$ which reflects the assumption that at time T the true disease status is observed.

2.4.1. Dynamic Rule In the two group case, a classical alternative to the simple rule $\widehat{U}(t) = \mathsf{argmax}_{g=0,1}\, \widehat{\mathcal{P}}_g(t)$ which allows us to modify the predictive accuracy measures (such as sensitivity or specificity) is to determine a cutoff value c and then to use the classification rule

$$\begin{aligned} \widehat{U}(t) &= 0, \quad \text{if } \widehat{P}_{\mathsf{D}}(t) \leq c, \\ \widehat{U}(t) &= 1, \quad \text{if } \widehat{P}_{\mathsf{D}}(t) > c. \end{aligned} \tag{7}$$

Then, indeed, at each visit, classification is determined.

In context of LoDA, this rule was basically considered, e.g., by Brant *et al.* [2] who additionally modified it to reflect a situation where it was more important to detect as early as possible patients in the disease group (as clinical action followed such classification) whereas it was not harmful for patients who truly belonged to the non-disease groups to remain under follow-up. This leads to the following classification rule. For $j = 1, \ldots, n-1$:

$$\begin{aligned} \widehat{U}(t_j) &= 1, \quad \text{if } \widehat{P}_{\mathsf{D}}(t_j) > c. \\ \widehat{U}(t_j) &= \mathsf{N.A.}, \quad \text{if } \widehat{P}_{\mathsf{D}}(t_j) \leq c. \end{aligned} \tag{8}$$

For the last visit at $t = t_n$, the classical rule (7) is applied and only then patient may be classified as non-diseased. In the following, we will refer to the rule (8) as dynLoDA.

The slightly modified version of the basic rule (7) is,

$$\begin{aligned} \widehat{U}(t) &= 0, \quad \text{if } \widehat{P}_{\mathsf{D}}(t) < 1 - c, \\ \widehat{U}(t) &= 1, \quad \text{if } \widehat{P}_{\mathsf{D}}(t) > c, \\ \widehat{U}(t) &= \mathsf{N.A.}, \quad \text{otherwise}, \end{aligned} \tag{9}$$

where the cutoff $c > 0.5$. Note that only those patients where the group pertinence is likely enough are classified.

2.4.2. Dynamic Rule based on two consecutive high probabilities In some situations, getting a single high probability of being in a disease group is not seen as strong evidence to allocate the individual to the disease group, and two consecutive high probabilities may be preferable in order to determine persistent high risk patients (see for example Reddy *et al.* [20]).

That is, under this rule, classification at time t_j, $j = 2, \ldots, n-1$ that directly generalizes the rule (8) is given by

$$\begin{aligned} \widehat{U}(t_j) &= 1, \qquad \text{if } \widehat{P}_{\mathsf{D}}(t_j) > c \ \ \& \ \ \widehat{P}_{\mathsf{D}}(t_{j-1}) > c, \\ \widehat{U}(t_j) &= \mathsf{N.A.}, \quad \text{otherwise.} \end{aligned} \tag{10}$$

At the first visit, the patient's status is always unclassified, i.e., $\widehat{U}(t_1) = \mathsf{N.A.}$ irrespective of available data, at the last visit at time $t = t_n$, the unclassified case is replaced by classification into the non-disease group.

By using this rule, classification is likely to take longer (leading to larger prediction times) than when using the classification rule based on a single group probability. Nevertheless, it may be necessary in situations where biomarkers used for prediction show large variability.

A further extension of this rule consists of identifying a change in the disease group probability between two consecutive visits. For example, if the probability of belonging to the disease group increases by more than a quantity, k, then this could be indicative of a worsening condition, and the patient could be classified into the disease group.

2.4.3. Fixed stopping time Hansen *et al.* [21] and Lukasiewicz *et al.* [22] consider scenarios where the aim is to identify non-responders (diseased patients) at an early date (before outcome is confirmed) while maintaining good levels of classification accuracy. They explore a number of fixed visit times and compare classification accuracies obtained by only considering data up until a pre-specified time, $T_{pred} < T$. That is, only one prediction is performed at time T_{pred} and the patient is classified into the disease group if $\widehat{\mathcal{P}}_D(T_{pred}) > c$, otherwise, the patient is classified as disease-free.

2.4.4. Selecting the cutoff There are a number of ways to choose the cutoff, c, of the classification rules outlined above. A typical measure is to use the cutoff associated with the point on the Receiver Operator Characteristic (ROC) curve nearest to the top left corner. Hence, the cutoff is chosen to minimise $d^2 = (1 - \mathsf{Sensitivity})^2 + (1 - \mathsf{Specificity})^2$. The ROC curve itself and related quantities are typically calculated using a suitable cross-validation procedure using the historical data with known group membership. An alternative method would be to use the Youden index [23]. In this method, the distance between sensitivity plus specificity and the chance line or diagonal line (i.e. $\mathsf{Sensitivity} + \mathsf{Specificity} - 1$) is maximised.

Further options include choosing the cutoff that maximises an alternative measure of classification accuracy such as the Probability of correct classification (PCC), positive predictive value (PPV) or negative predictive value (NPV). Variations on any of these measures are possible where conditions are made on some of the accuracy measures (e.g. Youden Index subject to sensitivity of at least 0.8), although such methods require some specification from the investigator. A comparison of eleven different cutoff selection options is given by Freeman and Moisen[24].

3. Dynamic Credible Intervals Prediction

All classification schemes outlined in Section 2.4 are based on point estimates of the disease group probability $\mathcal{P}_{\mathsf{D}}(t; \boldsymbol{\psi}, \boldsymbol{\theta})$. In a Bayesian context, the estimated posterior mean (6) is used here. In a frequentist context, one would use $\mathcal{P}_{\mathsf{D}}(t; \widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\theta}})$ instead, where $\widehat{\boldsymbol{\psi}}$ and $\widehat{\boldsymbol{\theta}}$ would be estimates of the model parameters based on the historical data. However, none of the outlined classification rules take into account uncertainty of the model parameter estimators. In a Bayesian context, the group membership probability $\mathcal{P}_{\mathsf{D}}(t; \boldsymbol{\psi}, \boldsymbol{\theta}) = \mathsf{P}\big(U = \mathsf{D} \,\big|\, \mathbb{Y}(t), \mathcal{C}(t); \boldsymbol{\psi}, \boldsymbol{\theta}\big)$ is just a derived model parameter and its uncertainty is reflected by variability of its posterior distribution. This posterior variability not only accounts for uncertainty attributed to estimation of the model parameters using the historical data but also the amount of information which is borne by the longitudinal and covariate history $\mathbb{Y}(t)$ and $\mathcal{C}(t)$ available at time t for a (new) patient

Statistics in Medicine

for whom the prediction is made. See bottom panels of Figures 2 and 3 for illustration of different variability in the posterior distribution of the group membership probabilities for different patients and also over time for one patient.

The variability of the posterior distribution of a quantity estimated in a Bayesian model can be summarized by its $(1-\alpha)100\%$ credible interval. The $(1-\alpha)100\%$ credible interval for $\mathcal{P}_{\mathsf{D}}(t;\ \boldsymbol{\psi}, \boldsymbol{\theta})$ is $\left(\mathcal{P}_{\mathsf{D}}^{LOW}(t),\ \mathcal{P}_{\mathsf{D}}^{UPP}(t)\right)$ if

$$\mathsf{P}\left[\mathcal{P}_{\mathsf{D}}(t;\ \boldsymbol{\psi}, \boldsymbol{\theta}) \in \left(\mathcal{P}_{\mathsf{D}}^{LOW}(t),\ \mathcal{P}_{\mathsf{D}}^{UPP}(t)\right) \,\middle|\, \mathcal{Y}\right] \ = \ 1-\alpha. \tag{11}$$

Condition (11) does not uniquely determine the credible interval as it only states that the interval captures $1-\alpha$ of the probability mass of the posterior distribution of $\mathcal{P}_{\mathsf{D}}(t;\ \boldsymbol{\psi}, \boldsymbol{\theta})$. If the corresponding posterior distribution is unimodal (as is mostly the case in our setting) then the shortest credible interval is called the *highest posterior density (HPD)* credible interval, (see Section 5.5 in Robert [25]). The HPD interval can easily be calculated once the MCMC sample $\left(\boldsymbol{\psi}^{(m)}, \boldsymbol{\theta}^{(m)}\right)$, $m = 1, \ldots, M$, from the posterior distribution $\left[\boldsymbol{\psi}, \boldsymbol{\theta} \,\middle|\, \mathcal{Y}\right]$ of the primary model parameters is available. In the following, let $\left(\mathcal{P}_{\mathsf{D}}^{LOW}(t),\ \mathcal{P}_{\mathsf{D}}^{UPP}(t)\right)$ denote the $(1-\alpha)100\%$ HPD credible interval for $\mathcal{P}_{\mathsf{D}}(t;\ \boldsymbol{\psi}, \boldsymbol{\theta})$.

In summary, the (HPD) credible interval captures the level of uncertainty of the corresponding group membership probability and provides an indication of the level of confidence one can have in the classification given. We propose a new classification scheme based on the use of credible intervals within the framework of dynamic LoDA. The idea is to incorporate the additional information provided by the credible intervals in a dynamic classification rule to create a procedure that achieves better values of prediction accuracy than procedures based purely on the point estimates.

The newly proposed *dynamic credible interval* (dynCI) classification rule to predict, at time t, the status of a (new) patient, again based on a cutoff value $c \in (0,\ 1)$ is as follows:

Steps of the *Dynamic Credible Interval* Classification Rule:

1. Calculate the $(1-\alpha)100\%$ (e.g., 95%) credible interval $\left(\mathcal{P}_{\mathsf{D}}^{LOW}(t_j),\ \mathcal{P}_{\mathsf{D}}^{UPP}(t_j)\right)$ for the disease group membership probability $\mathcal{P}_{\mathsf{D}}(t_j;\ \boldsymbol{\psi}, \boldsymbol{\theta})$.
2. If $\mathcal{P}_{\mathsf{D}}^{LOW}(t_j) > c$ assign patient to the disease group ($\widehat{U}(t_j) = 1$).
3. If $\mathcal{P}_{\mathsf{D}}^{UPP}(t_j) < c$ assign patient to the disease free group ($\widehat{U}(t_j) = 0$).
4. If $\mathcal{P}_{\mathsf{D}}^{LOW}(t_j) \ \leq \ c \ \leq \ \mathcal{P}_{\mathsf{D}}^{UPP}(t_j)$ the patient is left unclassified ($\widehat{U}(t_j) = \mathsf{N.A.}$).
5. If the patient remains unclassified, then when new information becomes available for the patient, update the probabilities calculated in Step 1, and follow steps 2-5.
6. Continue until patient is classified or observations from all visits are used.

The primary motivation for specification of steps 2. and 3. is to apply the basic rule (7) but only if we are sufficiently certain as expressed by the posterior probability of at least $1-\alpha$ that the disease group membership probability is higher then the cutoff c (step 2.) or lower than the cutoff c. The prediction time T_{pred} for a patient with (scheduled) visits at times $0 \leq t_1 < \cdots < t_n < T$ is then

$$T_{pred} \ = \ \min\left\{t_j :\ \mathcal{P}_{\mathsf{D}}^{LOW}(t_j) > c \ \text{ or } \ \mathcal{P}_{\mathsf{D}}^{UPP}(t_j) < c,\ j = 1, \ldots, n\right\}.$$

If the clinical interest is in only one particular group, then classification could focus on the classification conditions of that group, and patients classified in the other group could be 'treated' in the same way as the unclassified patients who remain under observation. For example, in the SANAD example presented further in this paper, the allocation scheme has been specifically designed to identify so called refractory patients (as opposed to the remission patients). When a patient is classified as refractory, prediction stops for this patient so that alternative treatment options can be considered. On the other hand, a patient who is classified as being in remission (by step 3 of the rule) remains under observation until they

achieve remission or their five years ($T = 5$ years) since diagnosis is up. So in our case, steps 3 and 4 are combined until the final visit for each patient, and any patient not classified as refractory is treated as unclassified (ie $\widehat{U}(t) = \text{N.A.}$). At the final visit considered for each patient (the visit before true status is determined) the patient could be classified as refractory, remission, or, if the status is still unclear, unclassified.

3.1. Benefits of using credible intervals

One potential benefit of the dynCI scheme is that some of the false positives and false negatives generated with other schemes may now be left unclassified. In our application, three groups of patients are generated under the dynCI scheme, a group of patients confidently classified as refractory, a second group of patients we are confident will achieve remission and a third group of patients we are unsure how to classify. We may be able to identify some patients about whom we are confident of their classification at an early stage, allowing alternative treatments to be considered much earlier to those patients whilst waiting longer for patients we are less confident about their group membership.

3.2. How to treat the unclassified group

An interesting question is what should be done with the unclassified patients. The answer to this depends on what the clinical scenario is. In our SANAD example, the unclassified group and the remission group are equally treated up until the final clinic visit for each patient (when a distinction is made between remission and unclassified patients), as clinicians would simply continue to observe them at regular intervals. One might consider merging the remission group and the unclassified group since the main aim in this case is to be more confident in prediction of patients with refractory epilepsy.

In a screening setting the groups could be used to influence personalised screening schedules. For example, in a study of diabetic retinopathy, the patients predicted as being likely to develop sight threatening diabetic retinopathy could be assigned to more frequent follow up appointments. Those patients we were confident were low risk could be assigned less frequent appointments than currently allocated, whilst the unclassified group could remain on the existing follow up schedule (annual screening intervals).

A different clinical application relates to diagnostic tests. For example, a cheap and non-invasive diagnostic test that uses data from a number of urine samples taken over time may be offered as an alternative to a more accurate but invasive and expensive diagnostic test (e.g., biopsy). In this case, patients identified by the LoDA procedure as belonging to the disease group could be assigned further treatment, whilst patients from the unclassified group would be individuals that the clinicians may wish to offer the invasive test.

4. Epilepsy Example

In this section we explore the merits of the dynCI allocation scheme using data from the SANAD study. In total, 1752 patients were considered of which 1577 were observed to achieve remission within five years of diagnosis, whilst 175 were not, and hence categorised as refractory. This data has been previously analysed in the context of LoDA by Hughes *et al.* [10] and we consider the same model here. We fit a MGLMM consisting of three longitudinal biomarkers: a binary marker indicating whether the patient had suffered seizures since their last visit, a continuous marker detailing the number of seizures since the last visit (transformed by $log(count + 1)$), and a Poisson distributed marker indicating the number of adverse events experienced since the previous clinic visit. The analysis here considers 20 fewer patients than were analysed in Hughes *et al.* [10] who were later discovered not to have epilepsy and their seizures were related to other causes. Nevertheless, the sensitivities and specificities obtained here are very similar to those reported in Hughes *et al.* [10] so we conclude the removal of these 20 patients does not affect the overall accuracy of the LoDA methods.

We use as explanatory covariates the time since the last follow up (as the visit schedule is irregular), gender, age at diagnosis of epilepsy, time since diagnosis, type of epilepsy (a binary indicator as to whether the patient has generalized

Statistics
in Medicine

epilepsy or not) and a binary covariate indicating whether or not the diagnosis occurred before the 6th June 2001. This variable was included to reflect the fact that a new drug was added to the trial on this date. To model the random effects we allow a random intercept and assume that the random effects follow a 2 component normal mixture distribution. More information on the SANAD data used in this analysis can be found in Hughes *et al.* [10] and Marson *et al.* [11, 12].

We present here the results of a leave-one-out cross validation study. For each patient, the MGLMM is fit using all the other patients but excluding the patient for whom prediction is to be made, and then LoDA is used to determine the group membership probabilities at each clinic visit. The cross validation study was set up in this manner to provide a clear picture of the predictions, and their associated variability for each individual patient. Calculations were performed in R [26], and the MGLMMs and LoDA were performed using the package mixAK [27].

4.1. Dynamic Rule using credible intervals

In this section we compare the prediction accuracy between the dynLoDA and dynCI classification schemes, and investigate the effect that level of credibility has on the accuracy. Specifically we consider 99%, 95%, 90% and 50% credible intervals within the dynCI scheme.

Table 1 shows the classification results using the dynLoDA scheme (panel (a)) and the dynCI scheme for various levels of credibility (panels (b)–(e)). Note that the total number of patients, 1685, is less than the 1752 patients considered. The 'missing' 67 patients (3.8%) had only one recorded visit, at which their status was confirmed, and therefore prediction was not possible for these patients. These patients could be used for model fitting however. The optimal cutoff 0.83 is determined by the d^2 rule of Section 2.4. With the exception of the 50% HPD, using the dynCI scheme reduces the number of both false positives and false negatives. This is because a percentage of patients wrongly classified under the dynLoDA scheme are left unclassified when using the dynCI scheme.

Table 2 summarises the predictive accuracy of the dynCI schemes with various levels of credibility. At this point, we consider the same cutoff for each scheme in order to compare directly the effect of using the dynCI scheme. Specificity, Sensitivity and PCC of the classified patients increase slightly when using the dynCI scheme. The most noticeable improvement is in the PPV, which increases from 0.56 in the dynLoDA scheme to 0.69 in our 99% dynCI scheme. Lead and prediction times were calculated for patients correctly identified as refractory and the means are provided in Table 2. Predictions tend to occur later with the dynCI scheme than with the dynLoDA scheme. As can be expected, the higher the level of credibility the longer it takes to classify a patient. Nevertheless, the cost in delay for using dynCI is less than 4 months in lead time, and there is still a good time gain since accurate predictions are made before true outcome is clinically confirmed (lead times between 565 and 661 days or between 18 and 22 months).

The overall levels of specificity and PCC decrease slightly using the 99%dynCI scheme and the sensitivity drops from 0.93 to 0.86. Comparison of panels (a) and (b) of Table 1 reveals that this drop in sensitivity was due to 16 truly refractory patients remaining unclassified, of whom 12 had been correctly identified using the dynLoDA scheme. Since a higher proportion of correctly identified refractory cases than misclassified refractory cases were moved to the unclassified category the overall sensitivity inevitably drops. However, we would argue that the benefit of the dynCI method is the greater level of confidence obtained for the classified patients (with an increase in PPV of over 10% (from 0.56 to 0.69)). In particular, 59 of the false positive cases are now unclassified the clinicians can be more confident in considering further treatment for patients classified as refractory. There is an increase in PPV using the dynCI scheme (except for the 50% HPD), since a higher percentage of patients predicted as refractory are in fact truly refractory, increasing the usefulness of the classification. The gain in PPV, and in other predictive accuracy measures, decreases as the level of credibility decreases. This is to be expected, as the narrowing of the credible interval can be thought of as a convergence towards the mean group membership probability. Hence we recommend the use of wide credible intervals in our setting.

A small number of patients are in the unclassified group with between 1% and 5% of patients not classified depending on the credibility level (Table 2). As expected, as the level of the credible intervals decreases, the credible intervals narrow and the number of patients in the unclassified group is reduced. The 99% dynCI scheme increases the specificity from

0.92 (obtained via the dynLoDA rule) to over 0.95 (only those patients who are classified are used in the calculation), whilst the sensitivity increases slightly (from 0.93 to 0.95).

TABLE 1 ABOUT HERE.

TABLE 2 ABOUT HERE.

4.2. Why use Credible Intervals?

The schemes in Section 2.4 only used, in a Bayesian setting, the posterior mean group membership probabilities to make a decision on a patient's group membership. We suggest that, due to posterior variability of the group membership probabilities, there may be substantial heterogeneity between patients in terms of the credibility of their estimated probabilities (their posterior means). This can be seen by considering two refractory patients whose longitudinal data is shown in Figure 1 and probability of being in the refractory group over time is shown for patient (a) in Figure 2, and for patient (b) in Figure 3. Histograms reflecting the posterior distribution of the probability of being in the refractory group (obtained from 10000 realisations of the MCMC sample) are also shown at each clinic visit. The final visit for patient (a), and the last five visits for patient (b) are not shown as the histograms do not differ much from the last visit shown.

Patient (a) experienced seizures at each clinic visit. As a result, even very close to the initial diagnosis the probability that he would ultimately be in the refractory group is around 0.4. When the number of seizures experienced increased at the third and fourth visits, his probability of belonging to the refractory group rose significantly to around 0.8. Since he

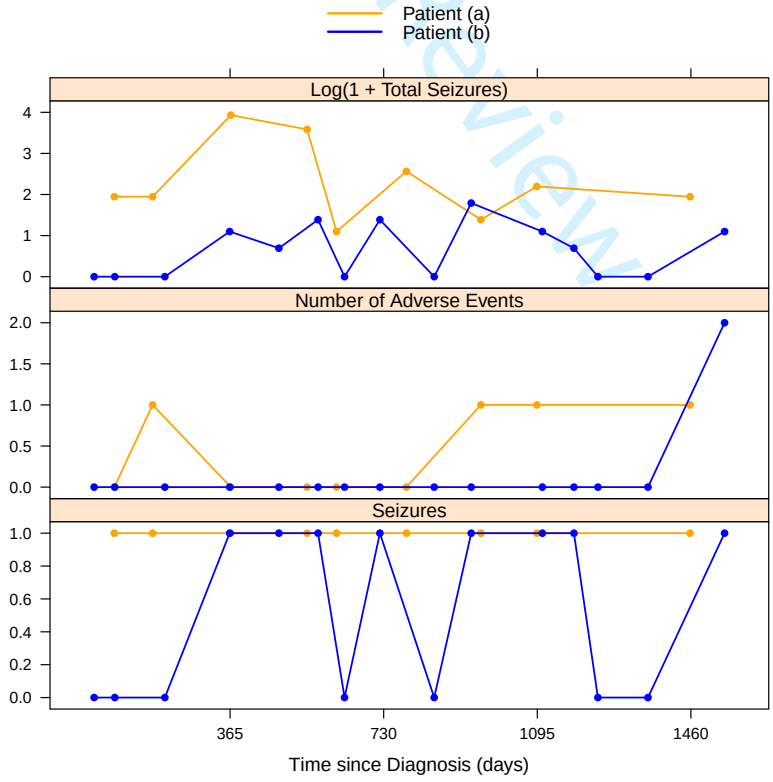


Figure 1. The longitudinal observations of two patients from SANAD. Patient (a) was a 17 year old male and Patient (b) was an 8 year old female. Both had generalised epilepsy diagnosed before 6th June 2001.

Statistics
in Medicine

kept having seizures, and his condition did not seem to be improving, it was very likely that he would ultimately be in the refractory group, which is reflected in the fact that the credibility bands are reasonably narrow for patient (a) and the histograms are all tightly clustered around the mean group membership probability.

Patient (b) had much more variety in her longitudinal observations. At some visits she had experienced seizures, whilst at others she had not. Consequently, her probability of being in the refractory group changed more noticeably over time, in contrast to the steady increase of patient (a). Her initial period free of seizures resulted in low probabilities that she would ultimately belong to the refractory group. At her fourth visit she experienced seizures (but only two, and no adverse events experienced). However, our model takes into account the full longitudinal history of the patient and not just her current information, so the fact that she had been seizure free until this point resulted in her probability of being refractory being relatively low. Once she had experienced seizures at her subsequent two visits the probability of being refractory continued to rise, but the uncertainty for this patient, due to her initial good response was shown in the wide credible intervals and wide histograms (high posterior variability) for visits 5–9. Only after she had been observed to have seizures for a longer period does her credible interval begin to narrow (visit 10), allowing her to be correctly identified as refractory.

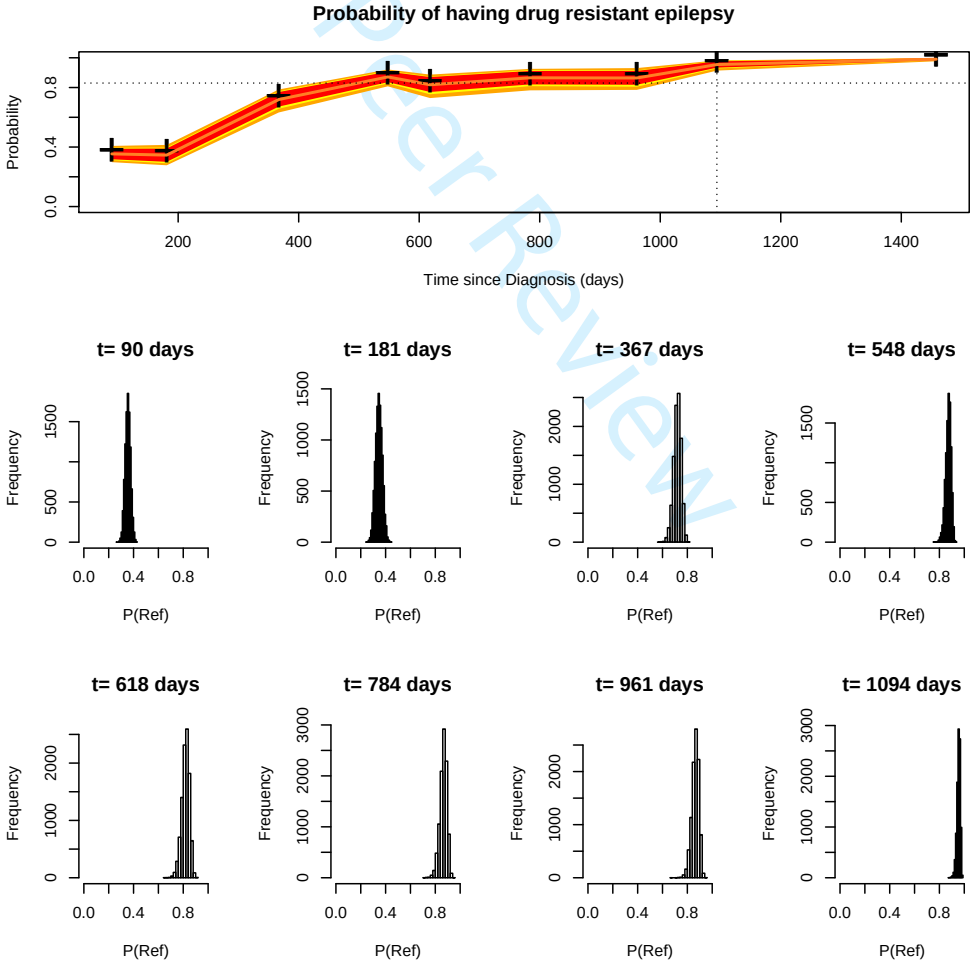


Figure 2. Marginal group membership probabilities over time for Patient (a) (top panel). The patient's probability of being refractory with 99%, 95% and 90% HPD intervals are represented by the orange, yellow and red areas respectively. Histograms estimating the posterior distribution of the probability of being in the refractory group are shown for each clinic visit below the top panel. The dotted vertical line denotes the time at which the patient was classified as refractory using the dynCI scheme with 99% credible intervals. The dotted horizontal line shows the required cutoff of 0.83.

Prepared using simauth.cls

5. Comparison of Classification Rules

Sections 2.4 and 3 describe four different allocation schemes, dynCI, dynLoDA, dynLoDA with two consecutive probabilities above a cutoff and a fixed time prediction scheme. Five ways of selecting the cutoff value are discussed, d^2, Youden's index and maximising PCC, PPV and NPV. We compare the allocation schemes under each choice of cutoff.

For the fixed time prediction scheme we consider making a prediction of a patient's status at $t_{pred} = 1, 2, 3, 4$ years. At each time point, t_{pred}, we focus on patients whose true status has not been determined at that point (we exclude patients who achieved remission before t_{pred}). Prediction is made using all the available data for each patient up until t_{pred}. For all other allocation schemes the classification is made dynamically following the rules outlined in Section 2.4. A 99% credible interval was used for the dynCI scheme since this gave the best predictive accuracy (Table 2).

TABLE 3 ABOUT HERE.

Table 3 shows summaries of the classification accuracy using the allocation schemes. Prediction based solely on a fixed time point (one, two, three or four years post diagnosis) yields worse prediction accuracy than when allowing a dynamic

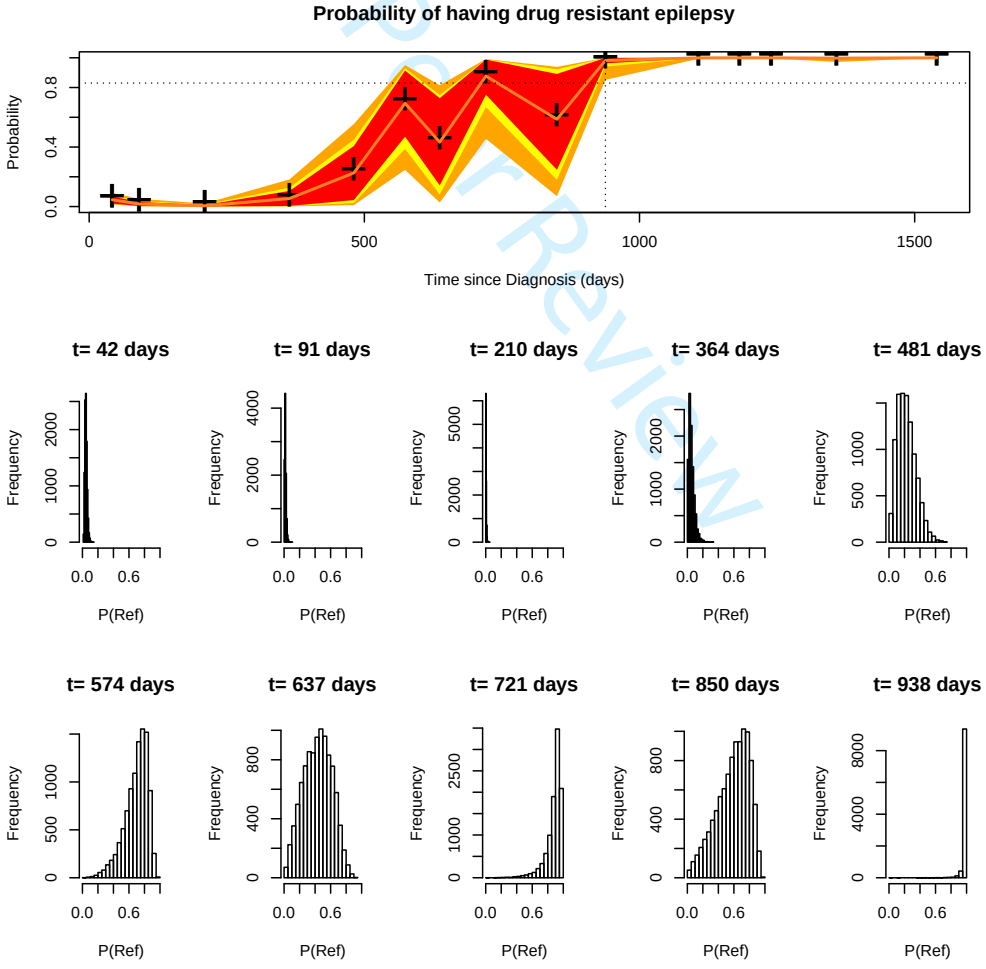


Figure 3. Marginal group membership probabilities over time for Patient (b) (top panel). The patient's probability of being refractory with 99%, 95% and 90% HPD intervals are represented by the orange, yellow and red areas respectively. Histograms showing the posterior distribution of the probability of being in the refractory group are shown for each clinic visit below the top panel. The dotted vertical line denotes the time at which the patient was classified as refractory using the dynCI scheme with 99% credible intervals. The dotted horizontal line shows the required cutoff of 0.83.

Statistics
in Medicine

prediction scheme. This result indicates the existence of heterogeneity amongst patients. Some patients are easy to identify relatively early, whilst others take longer to determine their status. A dynamic scheme, which predicts patients as refractory at the time we are confident about the prediction, allows flexible stopping rules based on individual patient data.

The cost of maximising PPV and NPV is a significant reduction in sensitivity and specificity respectively. Whether this would be desirable depends on the clinical situation. With the dynLoDA scheme, selecting the cutoff to maximise PCC only yields a 3% increase in PCC compared to the d^2 and Youden methods, but the cost is a 18% drop in sensitivity. When using the dynCI scheme, there is a slight increment in sensitivity, specificity and PCC, as well as a noticeable improvement in the change in PPV (from 0.56 to 0.69 using the d^2 rule) compared to the dynLoDA scheme. The cutoff could be chosen to maximise the PPV. However, with the dynCI scheme, this would yield an 11% increase in PPV (0.69-0.8) but at the cost of an 11% drop in sensitivity and would require an extra 190 days on average (approx 6 months) to correctly classify refractory patients. The dynCI scheme could be seen as a way of improving PPV without sacrificing good levels of sensitivity, specificity and PCC. The cost to this, in comparison to the dynLoDA rule, as mentioned in Section 4.1, is a delay in time of classification (on average 4 months as the allocation scheme is more conservative).

Figure 4 shows the performance of each classification rule over a range of possible cutoff values. The rules based on a fixed prediction time perform substantially worse than those which allow classification to be made dynamically. The dynLoDA scheme and dynLoDA scheme based on 2 consecutive high probabilities perform broadly similarly, although Table 3 shows that using the d^2 optimal cutoff, the dynLoDA scheme is slightly better in terms of predictive accuracy. The dynCI scheme with 99% credibility interval shows the best performance.

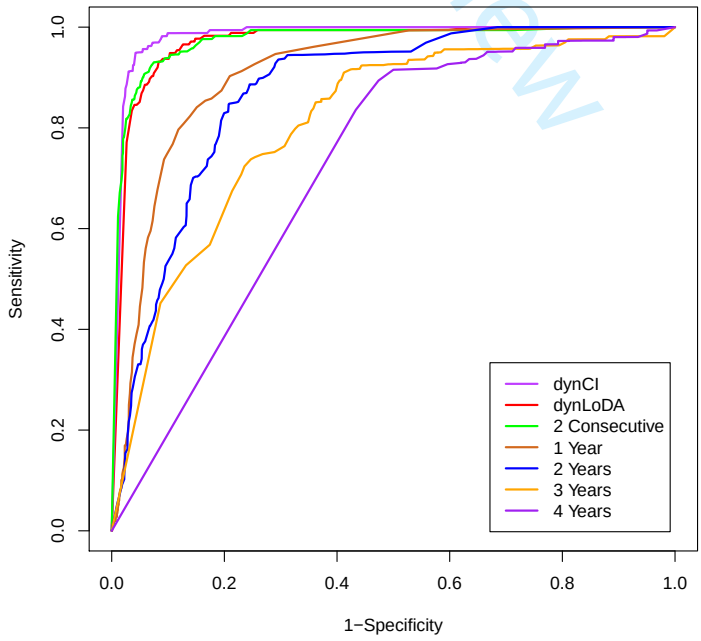


Figure 4. Receiver Operating Characteristic curves for each classification rule. The dynCI results are based on the classified patients with a 99% credibility interval.

Prepared using simauth.cls

6. Discussion

In this paper we propose the use of credible intervals for group membership probabilities to improve classification in a dynamic longitudinal discriminant analysis. The idea is to account for the variability in the level of uncertainty of the group membership probabilities that exists between individuals. The approach proposed here is both dynamic and personalised. It is dynamic since classification is updated each time new data becomes available and personalised because an individual patient's characteristics, both baseline and longitudinal data, but also individual levels of uncertainty of group membership are taken into account to aid classification. The dynCI model can be fit using the mixAK [27] package in R. We attach as supplementary material a simulated data set and also an R script to fit the MGLMMs and run the LoDA.

We apply our approach to clinical data from the SANAD study to identify in advance patients who will not achieve remission from seizures within 5 years from diagnosis. We show that the use of credible intervals in an allocation scheme can reduce the number of false positives, and so increase the PPV of the prognostic rule. Additionally, we have also shown that the sensitivity, specificity and PCC of classified patients can improve when using credible intervals. Dynamic prediction was considerably more accurate than when prediction was conducted for all patients at a fixed time point.

We demonstrate that the lead time is substantial with the dynCI classification. Therefore, such improvement in lead times could potentially allow clinicians to consider alternative treatment routes for some patients much earlier on than is currently practiced, which may improve patients' care. We also show that using credible intervals on the SANAD data, allows us to be more confident about the predictions given.

A 99% credible interval was used for the analysis in this paper as this gave the best improvement in prediction accuracy (Table 2) with only 5% of patients remaining unclassified. The choice of the level of credible interval is application specific and must balance the desire for increased accuracy in prediction with the aim of classifying most patients. How many patients a clinician is willing to leave unclassified depends to some extent on the severity of alternative treatments to be attempted for patients predicted to have the disease. In our epilepsy example, brain surgery might be considered for patients who are classified as refractory. Leaving more patients unclassified would be preferable to wrongly classifying them as refractory and putting them through unnecessary brain surgery. We recommend setting the level of the credible interval as high as possible whilst maintaining an application specific acceptable level of unclassified patients.

In this paper we have only considered classification into one of two prognostic groups. The classification scheme based on credible intervals is easily adaptable to a multiple group situation. In that case a given patient would be classified into group g at time t if the (estimated) posterior mean $\widehat{\mathcal{P}}_g(t)$, Eq. (6) is the largest of the posterior means of all the groups and also $\mathcal{P}_g^{LOW}(t)$, the lower limit of the credible interval for $\mathcal{P}_g(t;\ \boldsymbol{\psi},\ \boldsymbol{\theta})$ was greater than a given cutoff. Any patient not meeting this criteria for any of the groups would remain unclassified.

This work considers that once a patient is classified, their status is not revisited. Alternative statistical models may be considered to capture changes in status when individuals may move between the disease and non-disease groups over time (eg Multistate models, Commenges and Jacqmin-Gadda [28], Chapter 7).

We envisage that this work could be used to determine when a patient should next visit a clinic for a screening appointment. Initial work on personalised screening intervals has been done by Rizopoulos *et al.* [29]. We believe our dynCI rule could be applied so that the unclassified group maintain the currently advised schedule of visits, the disease free group could be allocated less frequent clinic visits and the disease group could be assigned more frequent clinic visits.

In this paper we have taken a longitudinal discriminant analysis approach to classify patient based on their longitudinal data. It would be possible to consider alternative models to classify patients dynamically. However each approach would require a reframing of the question of interest. For example one could consider survival based approaches such as landmarking [30] or joint modelling of longitudinal and time to event data [19]. In this case the focus would be on risk of achieving remission within the remaining time up until five years post diagnosis. Other alternatives could include machine learning methods for high dimensional data [31] and support vector machine learning methods for longitudinal data [32]. Although this paper considers longitudinal discriminant analysis, we believe that the credible intervals scheme could work

well in any dynamic prediction classification scheme including those mentioned above. In principle any scheme for which credible/confidence intervals are calculated around the probability in question could work well with our scheme.

This work has made used of a HPD credible interval in the allocation scheme. This makes the assumption that the posterior distribution for the group membership probability is unimodal. This can be checked on histograms of sampled values (see Figure 3) which can even be smoothed, e.g., by a suitable kernel density estimator (e.g., Silverman [33] Chapter 3) for better insight. It is our empirical experience that unimodality is mostly the case in analysis such as that presented in this paper. However, if the posterior distribution were severely multimodal, then our assumption may lead to inaccurate results since the highest posterior density region could be disjoint intervals (See Section 2.3 of Gelman et al. [34]). In such a case it may be preferable to use equal-tail credible intervals based on posterior quantiles.

Further work could be done to identify more complex stopping rules taking into account severity of alternative treatments, cost of treatment and possibly other factors. Such models may be more complicated but could allow clinicians to use more information when making decisions for individual patients.

Acknowledgements

The first, third, fourth and fifth author acknowledge support from the Medical Research Council (Research project MR/L010909/1). The third author is funded by a Post-Doctoral Fellowship from the National Institute of Health Research (PDF-2015-08-044). The fourth author is also grateful to the Clinical Eye Research Centre, St. Paul's Eye Unit, Royal Liverpool and Broadgreen University Hospitals NHS Trust for supporting this work. Support of the second author from the Interuniversity Attraction Poles Programme (IAP-network P7/06), Belgian Science Policy Office, is also gratefully acknowledged. We are grateful to Professor Anthony Marson for permission to use the SANAD data.

References

1. Fieuws S, Verbeke G, Maes B, Van Renterghem Y. Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics* 2008; **9**(3):419–431.
2. Brant LJ, Sheng SL, Morrell CH, Verbeke GN, Lesaffre E, Carter HB. Screening for prostate cancer by using random-effects models. *Journal of the Royal Statistical Society, Series A* 2003; **166**(1):51–62.
3. Tomasko L, Helms RW, Snapinn SM. A discriminant analysis extension to mixed models. *Statistics in Medicine* 1999; **18**(10):1249–1260.
4. Wernecke KD, Kalb G, Schink T, Wegner B. A mixed model approach to discriminant analysis with longitudinal data. *Biometrical Journal* 2004; **46**(2):246–254.
5. Lix LM, Sajobi TT. Discriminant analysis for repeated measures data: A review. *Frontiers in Psychology* 2010; **1**(146):1–9.
6. Kohlmann M, Held L, Grunert VP. Classification of therapy resistance based on longitudinal biomarker profiles. *Biometrical Journal* 2009; **51**(4):610–626.
7. Morrell CH, Brant LJ, Sheng S, Metter EJ. Screening for prostate cancer using multivariate mixed-effects models. *Journal of Applied Statistics* 2012; **39**(6):1151–1175.
8. Marshall G, De la Cruz-Mesía R, Quintana FA, Barón AE. Discriminant analysis for longitudinal data with multiple continuous responses and possibly missing data. *Biometrics* 2009; **65**(1):69–80.
9. Komárek A, Hansen BE, Kuiper EMM, van Buuren HR, Lesaffre E. Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statistics in Medicine* 2010; **29**(30):3267–3283.
10. Hughes DM, Komárek A, Czanner G, Garcia-Fiñana M. Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types. *Statistical Methods in Medical Research* 2016; :10.1177/0962280216674 496.
11. Marson AG, Al-Kharusi AM, Alwaidh M, Appleton R, Baker GA, Chadwick DW, Cramp C, Cockerell OC, Cooper PN, Doughty J, *et al.*. The SANAD study of effectiveness of valproate, lamotrigine, or topiramate for generalised and unclassifiable epilepsy: an unblinded randomised controlled trial. *The Lancet* 2007; **369**(9566):1016–1026.
12. Marson AG, Al-Kharusi AM, Alwaidh M, Appleton R, Baker GA, Chadwick DW, Cramp C, Cockerell OC, Cooper PN, Doughty J, *et al.*. The SANAD study of effectiveness of carbamazepine, gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: an unblinded randomised controlled trial. *The Lancet* 2007; **369**(9566):1000–1015.
13. Komárek A, Komárková L. Clustering for multivariate continuous and discrete longitudinal data. *The Annals of Applied Statistics* 2013; **7**(1):177–200.
14. Guglielmi A, Ieva F, Paganoni AM, Ruggeri F, Soriano J. Semiparametric Bayesian models for clustering and classification in the presence of unbalanced in-hospital survival. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2014; **63**(1):25–46.

15. Zhang X, Jeske DR, Li J, Wong V. A sequential logistic regression classifier based on mixed effects with applications to longitudinal data. *Computational Statistics & Data Analysis* 2016; **94**:238–249.
16. Horrocks J, van Den Heuvel MJ. Prediction of pregnancy: A joint model for longitudinal and binary data. *Bayesian Analysis* 2009; **4**(3):523–538.
17. Shah NH, Hipwell AE, Stepp SD, Chang CCH. Measures of discrimination for latent group-based trajectory models. *Journal of Applied Statistics* 2015; **42**(1):1–11.
18. Morrell CH, Sheng SL, Brant LJ. A comparative study of approaches for predicting prostate cancer from longitudinal data. *Communications in Statistics – Simulation and Computation* 2011; **40**(9):1494–1513.
19. Rizopoulos D. *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press, 2012.
20. Reddy T, Molenberghs G, Njagi EN, Aerts M. A novel approach to estimation of the time to biomarker threshold: applications to hiv. *Pharmaceutical statistics* 2016; **15**(6):541–549.
21. Hansen BE, Komárek A, Buster EHCJ, Steyerberg EW, Janssen HLA, Lesaffre E. Dynamic prediction of response to HBV-treatment using multivariate longitudinal profiles. *Statistical Models of Treatment Effects in Chronic Hepatitis B and C*, Hansen BE (ed.). chap. 2.4, Erasmus Universiteit: Rotterdam, 2010; 79–103.
22. Lukasiewicz E, Gorfine M, Neumann AU, Freedman LS. Combining longitudinal discriminant analysis and partial area under the roc curve to predict non-response to treatment for hepatitis c virus. *Statistical methods in medical research* 2010; .
23. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3**(1):32–35.
24. Freeman EA, Moisen GG. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling* 2008; **217**(1):48–58.
25. Robert CP. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementations*. Second edn., Springer Science+Business Media: New York, 2007.
26. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2016. URL http://www.R-project.org/.
27. Komárek A, Komárková L. Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data. *Journal of Statistical Software* 2014; **59**(12):1–38. URL http://www.jstatsoft.org/v59/i12/.
28. Commenges D, Jacqmin-Gadda H. *Dynamical Biostatistical Models*, vol. 86. CRC Press, 2015.
29. Rizopoulos D, Taylor JMG, Van Rosmalen J, Steyerberg EW, Takkenberg JJM. Personalized screening intervals for biomarkers using joint models for longitudinal and survival data. *Biostatistics* 2015; .
30. van Houwelingen H, Putter H. *Dynamic prediction in clinical survival analysis*. CRC Press, 2011.
31. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second edn., Springer-Verlag: New York, 2009.
32. Luts J, Molenberghs G, Verbeke G, Van Huffel S, Suykens JA. A mixed effects least squares support vector machine model for classification of longitudinal data. *Computational Statistics & Data Analysis* 2012; **56**(3):611–628.
33. Silverman BW. *Density estimation for statistics and data analysis*, vol. 26. CRC press, 1986.
34. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian data analysis*, vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

Statistics
in Medicine

Table 1. Classification for the marginal prediction with a cutoff of 0.83 using the dynLoDA scheme (a), and the dynCI prediction with a level of 99% (b), 95% (c), 90% (d) and 50% (e).

(a)

		Classification		
		Remission	Refractory	Total
True Status	Remission	1384	126	1510
	Refractory	12	163	175
	Total	1396	289	1685

(b) 99% HPD CI

		Classification			
		Remission	Refractory	Unclassified	Total
True Status	Remission	1368	67	75	1510
	Refractory	8	151	16	175
	Total	1376	218	91	1685

(c) 95% HPD CI

		Classification			
		Remission	Refractory	Unclassified	Total
True Status	Remission	1368	90	52	1510
	Refractory	9	155	11	175
	Total	1377	245	63	1685

(d) 90% HPD CI

		Classification			
		Remission	Refractory	Unclassified	Total
True Status	Remission	1368	98	44	1510
	Refractory	10	156	9	175
	Total	1378	254	53	1685

(e) 50% HPD CI

		Classification			
		Remission	Refractory	Unclassified	Total
True Status	Remission	1361	131	18	1510
	Refractory	12	162	1	175
	Total	1373	293	19	1685

Prepared using simauth.cls

Table 2. A comparison of the prediction accuracy using dynLoDA and dynCI schemes. The sensitivities and specificities highlighted in bold are calculated without considering the unclassified patients.

	dynLoDA	dynCI 99% HPDCI	dynCI 95% HPDCI	dynCI 90% HPDCI	dynCI 50% HPDCI
Cutoff	0.83	0.83	0.83	0.83	0.83
Sensitivity (Classified Data)	**0.93**	**0.95**	**0.95**	**0.94**	**0.93**
Sensitivity	0.93	0.86	0.89	0.89	0.93
Specificity (Classified Data)	**0.92**	**0.95**	**0.94**	**0.93**	**0.91**
Specificity	0.92	0.91	0.91	0.91	0.90
PCC (Classified Data)	**0.92**	**0.95**	**0.94**	**0.93**	**0.91**
PCC	0.92	0.90	0.90	0.90	0.90
AUC	0.97	0.98	0.98	0.97	0.97
PPV	**0.56**	**0.69**	**0.63**	**0.61**	**0.55**
NPV	0.99	0.99	0.99	0.99	0.99
Proportion Unclassified	0.00	0.05	0.04	0.03	0.01
Mean Lead Time (days)	675	565	595	614	661
Mean Prediction Time (days)	857	972	942	918	872

PCC = Probability of Correct Classification, AUC = Area Under Curve, PPV = Positive Predictive Value, NPV = Negative Predictive Value.

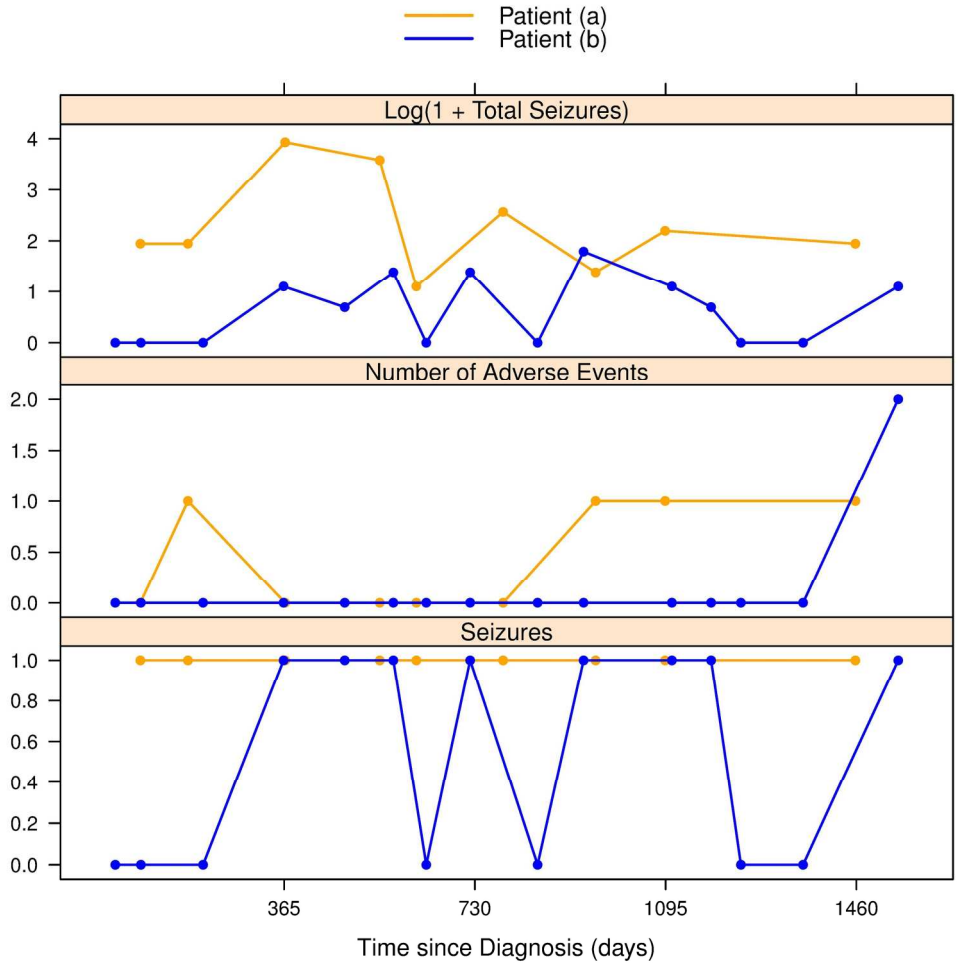
Statistics in Medicine

Table 3. Prediction accuracy using different allocation rules. Blank entries represent cases where no patients were classified as refractory. A 99% credibility interval was used for the dynCI scheme.

	d^2	Youden	Max PCC	Max PPV	Max NPV	d^2	Youden	Max PCC	Max PPV	Max NPV
			dynCI					**dynLoDA**		
Cutoff	0.83	0.83	0.99	0.99	0.01	0.83	0.83	0.99	0.99	0.01
Sensitivity	0.95	0.95	0.84	0.84	1.00	0.93	0.93	0.77	0.77	1.00
Specificity	0.95	0.95	0.98	0.98	0.11	0.92	0.92	0.97	0.97	0.12
PCC	0.95	0.95	0.97	0.97	0.24	0.92	0.92	0.95	0.95	0.21
AUC	0.98	0.95	0.95	0.95	0.95	0.97	0.97	0.97	0.97	0.97
PPV	0.69	0.69	0.80	0.80	0.16	0.56	0.56	0.77	0.77	0.12
NPV	0.99	0.99	0.98	0.98	1.00	0.99	0.99	0.97	0.97	1.00
Mean Lead Time (days)	565	565	378	378	1427	675	675	432	432	1427
Mean Prediction Time (days)	972	972	1162	1162	108	857	857	1106	1106	107
			Two Consecutive					**1 Year**		
Cutoff	0.58	0.49	0.97	0.99	0.01	0.10	0.10	1.00	0.54	0.01
Sensitivity	0.91	0.94	0.65	0.58	1.00	0.83	0.83	0.00	0.16	0.95
Specificity	0.89	0.86	0.98	0.98	0.33	0.73	0.73	1.00	0.96	0.53
PCC	0.89	0.87	0.94	0.94	0.41	0.74	0.74	0.90	0.88	0.57
AUC	0.96	0.96	0.96	0.96	0.96	0.83	0.83	0.83	0.83	0.83
PPV	0.52	0.47	0.79	0.81	0.16	0.26	0.26		0.32	0.19
NPV	0.99	0.99	0.96	0.95	1.00	0.97	0.97	0.90	0.91	0.99
Mean Lead Time (days)	703	799	386	337	1301	1250	1250		1194	1256
Mean Prediction Time (days)	831	735	1148	1202	233	280	280		312	277
			2 Years					**3 Years**		
Cutoff	0.31	0.21	0.91	0.95	0.01	0.87	0.77	0.87	0.99	0.16
Sensitivity	0.79	0.87	0.26	0.20	0.98	0.76	0.81	0.76	0.42	0.95
Specificity	0.67	0.62	0.95	0.97	0.29	0.67	0.62	0.67	0.86	0.31
PCC	0.69	0.67	0.80	0.80	0.44	0.71	0.70	0.71	0.67	0.58
AUC	0.80	0.80	0.80	0.80	0.80	0.75	0.75	0.75	0.75	0.75
PPV	0.40	0.38	0.58	0.62	0.28	0.63	0.61	0.63	0.69	0.50
NPV	0.92	0.95	0.82	0.81	0.98	0.79	0.81	0.79	0.67	0.90
Mean Lead Time (days)	929	937	887	866	943	588	593	588	546	615
Mean Prediction Time (days)	599	595	653	662	590	953	947	953	998	927
			4 Years							
Cutoff	0.99	0.93	0.91	0.96	0.02					
Sensitivity	0.76	0.89	0.90	0.85	0.99					
Specificity	0.57	0.50	0.49	0.53	0.06					
PCC	0.71	0.77	0.78	0.76	0.71					
AUC	0.70	0.70	0.70	0.70	0.70					
PPV	0.81	0.80	0.80	0.81	0.71					
NPV	0.51	0.67	0.69	0.61	0.80					
Mean Lead Time (days)	248	279	286	266	297					
Mean Prediction Time (days)	1309	1280	1272	1291	1261					

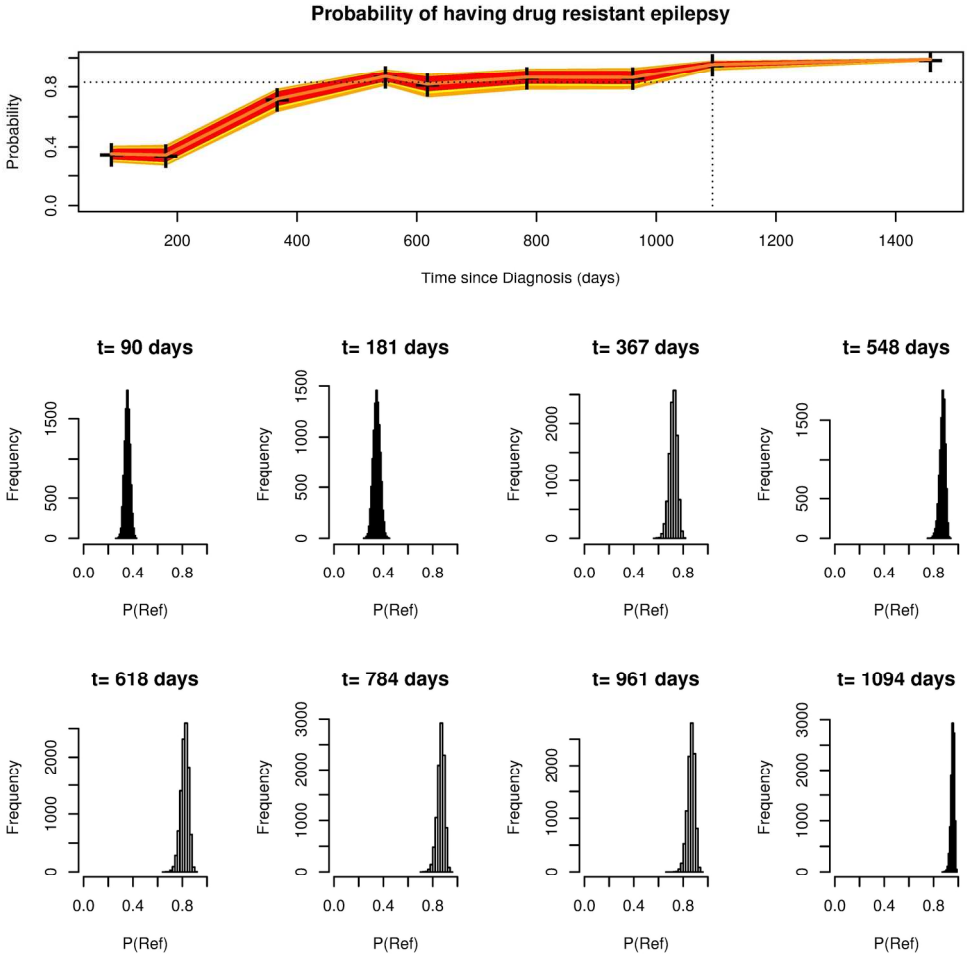
PCC = Probability of Correct Classification, AUC = Area Under Curve, PPV = Positive Predictive Value, NPV = Negative Predictive Value.

Prepared using simauth.cls



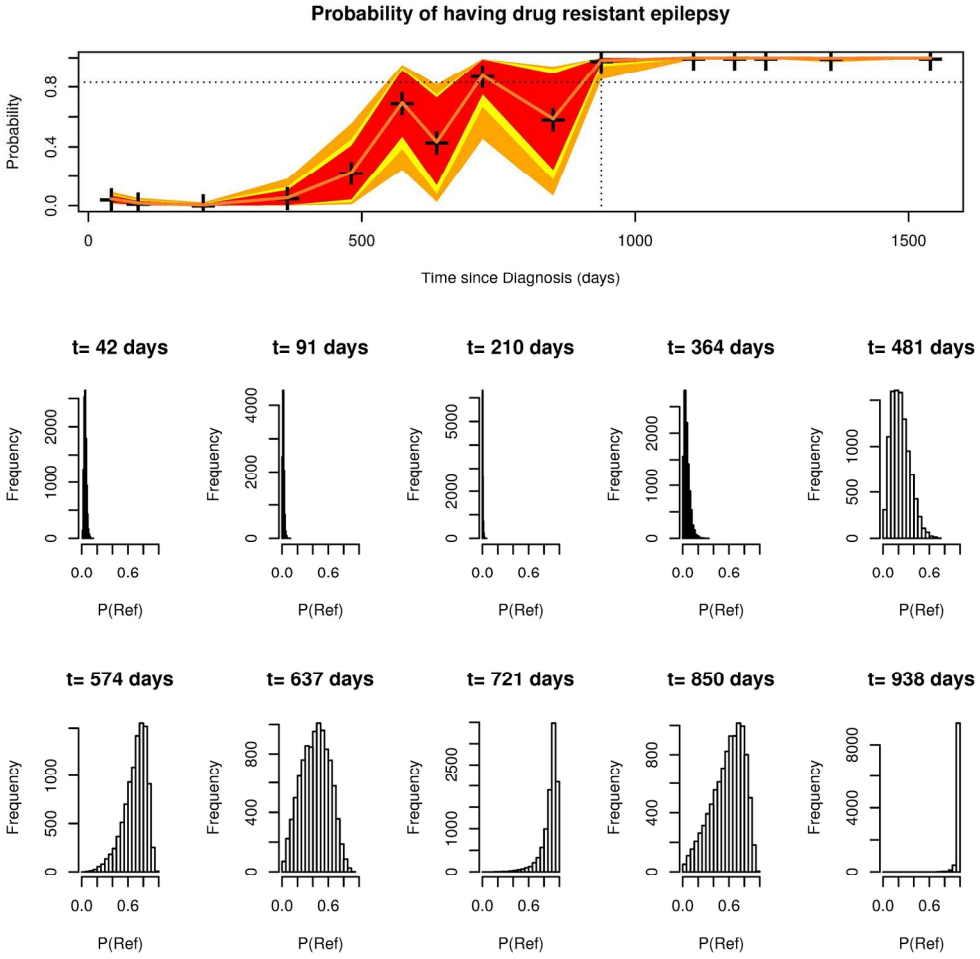
The longitudinal observations of two patients from SANAD. Patient (a) was a 17 year old male and Patient (b) was an 8 year old female. Both had generalised epilepsy diagnosed before 6th June 2001.

177x177mm (300 x 300 DPI)



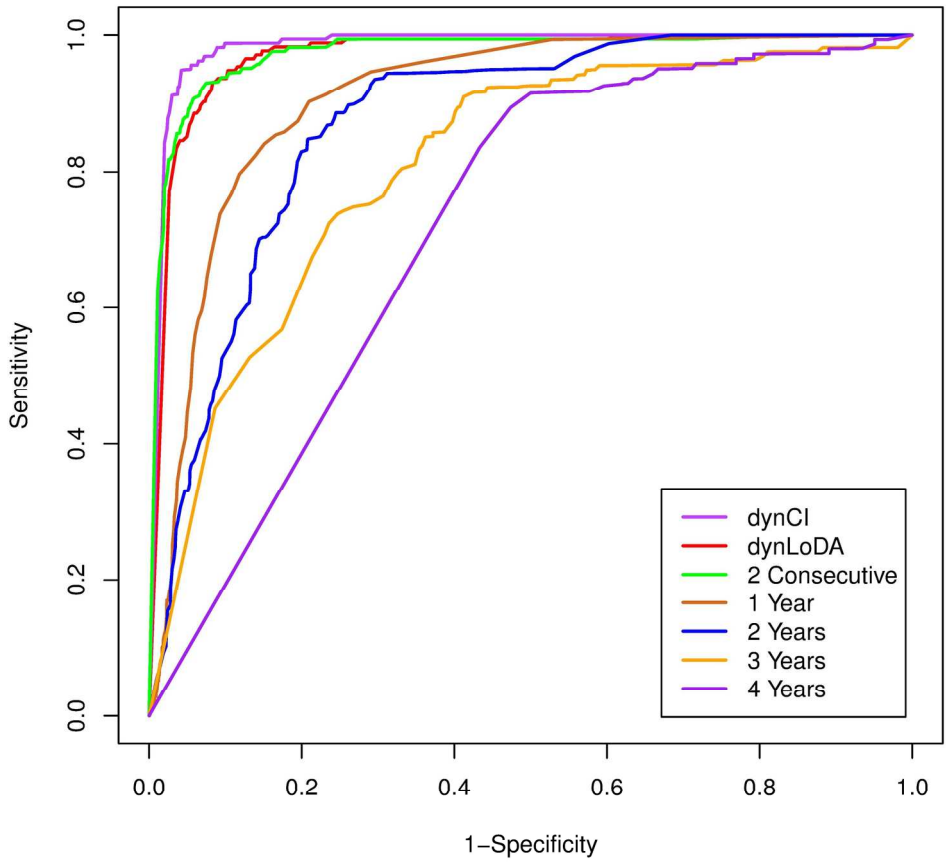
Marginal group membership probabilities over time for Patient (a) (top panel). The patient's probability of being refractory with 99%, 95% and 90% HPD intervals are represented by the orange, yellow and red areas respectively. Histograms estimating the posterior distribution of the probability of being in the refractory group are shown for each clinic visit below the top panel. The dotted vertical line denotes the time at which the patient was classified as refractory using the dynCI scheme with 99% credible intervals. The dotted horizontal line shows the required cutoff of 0.83.

177x177mm (300 x 300 DPI)



Marginal group membership probabilities over time for Patient (b) (top panel). The patient's probability of being refractory with 99%, 95% and 90% HPD intervals are represented by the orange, yellow and red areas respectively. Histograms showing the posterior distribution of the probability of being in the refractory group are shown for each clinic visit below the top panel. The dotted vertical line denotes the time at which the patient was classified as refractory using the dynCI scheme with 99% credible intervals. The dotted horizontal line shows the required cutoff of 0.83.

177x177mm (300 x 300 DPI)



Receiver Operating Characteristic curves for each classification rule. The dynCI results are based on the classified patients with a 99% credibility interval.

177x177mm (300 x 300 DPI)